Overflow Oscillations in Digital Filters

By P. M. EBERT, JAMES E. MAZO,
AND MICHAEL G. TAYLOR

(Manuscript received May 9, 1969)

The cascade and parallel realizations of an arbitrary digital filter are both formed using second order sections as building blocks. This simple recursive filter is commonly implemented using 2's complement arithmetic for the addition operation. Overflow can then occur at the adder and the resulting nonlinearity causes self-oscillations in the filter. The character of the resulting oscillations for the second order section are here analyzed in some detail. A simple necessary and sufficient condition on the feedback tap gains to insure stability, even with the presence of the nonlinearity, is given although for many desired designs this will be too restrictive. A second question studied is the effect of modifying the "arithmetic" in order to quench the oscillations. In particular it is proven that if the 2's complement adder is modified so that it "saturates" when overflow occurs, then no self-oscillations will be present.

I. INTRODUCTION

A digital filter using idealized operations can easily be designed to be stable.[1] Nevertheless, in actual implementations, the output of such a stable filter can display large oscillations even when no input is present.* A known cause of this phenomenon is the fact that the digital filter realization of the required addition operation can cause overflow, thereby creating a severe nonlinearity.† Our purpose here is twofold. The first is to give a somewhat detailed analysis of the character of the oscillations when the filter is a simple second order recursive section with two feedback taps. This unit is the fundamental building block for the cascade and the parallel realization of digital filters, and as such is worthy of some scrutiny.[2] A simple conclusion which one can draw from

* To the best of our knowledge, these oscillations were first observed and diagnosed by L. B. Jackson of Bell Telephone Laboratories.

† In the present work rounding errors in multiplication or storage are neglected and therefore so are the little-understood oscillations attendant upon these nonlinearities.

the analysis is that the design of many useful filters requires using values of feedhack coefficients such that the threat of oscillations is always present (witb 2's complement arithmetic). Optimum solutions that cope with this state of affairs are still unknown. Some recent proposals include observing when overflow at the adder is to occur and then taking appropriate action. Our second purpose, then, is to discuss tbe effectiveness of some of these ideas, and to give a proof that modifying 2's complement arithmetic so that the adder "saturates" is an effective way to eliminate the oscillations. Questions of how this nonlinearity will affect the desired outputs from a particular ensemhle of input signals are not yet answered however, and perhaps for some applications other solutions need he considered.

II. PROBLEM FORMULATION AND GENERAL DISCUSSION

As explained in the introduction, this paper deals primarily with the simple structure shown in Fig. 1. The outputs of the registers, which are storage elements with one unit of delay, are multiplied hy coefficients a and b respectively, fed hack, and "added" to the input in the accumulator. No round-off error is considered either in multiplication or storage, but overflow of the accumulator is not neglected. In other words, the accumulator will perform as a true adder if the sum of its inputs is in some range; otherwise a nonlinear hehavior is observed.

Figure 2 shows the instantaneous input-output characteristic $f(v)$ of the device motivated hy using 2's complement arithmetic. It is also important to note that there is no memory of the accumulator for past outputs; that is, the device is zeroed after the generation of each output.

If we let $x(t)$ be the input signal to the device, $y(t)$ the output, and

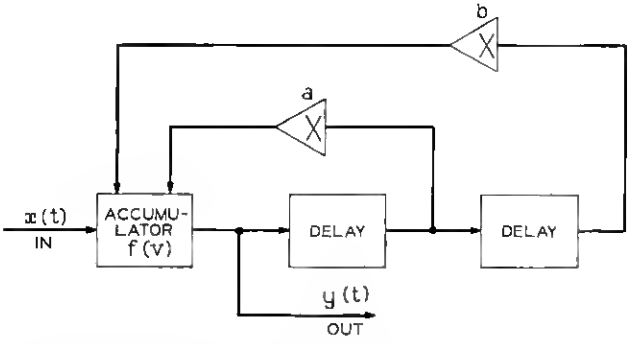

Fig. 1 — Basic configuration for the digital filter $\cdot y_{k+2} = f[ay_{k+1} + by_k + x_{k+2}]$.

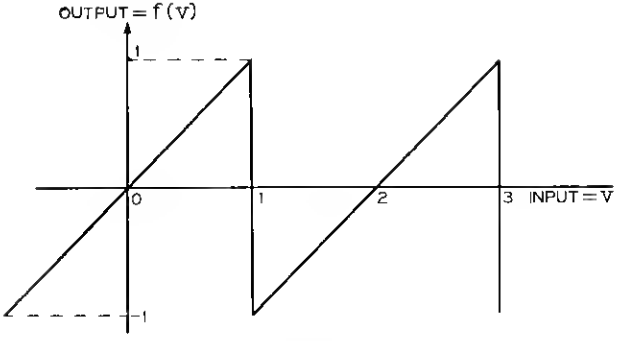

Fig. 2 — Instantaneous transfer function of the accumulator.

$f(\cdot)$ the nonlinear characteristic of the accumulator, we have the basic equation

$$y(t + 2) = f[ay(t + 1) + by(t) + x(t + 2)]. \tag{1}$$

We shall be concerned with the self-sustaining oscillations of the device that are observed even when no input is present [$x(t) = 0$], and when linear theory would predict the device to be stable.

By making this linear approximation $f(v) = v$, the linearized version of equation (1) becomes, with no driving term in the equation,

$$y(t + 2) - ay(t + 1) - by(t) = 0. \tag{2}$$

The roots of the characteristic equation for equation (2) are

$$\rho_{1,2} = \frac{a \pm (a^2 + 4b)^{\frac{1}{2}}}{2} \tag{3}$$

and the region of linear stability corresponds to the requirement that $| \rho_i | < 1$. This region is depicted as a subset of the a–b plane in Fig. 3. One has $| \rho_i | < 1$ if and only if one is within the large triangle shown in Fig. 3. For this situation any solution of (2) will damp out to zero after a sufficient period of time. Now note that (2) is not necessarily a valid reduction of (1) even when $x(t) = 0$. The output, by choice of f, has been assumed to be constrained to be less than unity, but this is not sufficient to guarantee that the argument of the function f is less than unity. For this to be the case we require

$$| ay(t + 1) + by(t) | < 1. \tag{4}$$

Since $| y(t) | < 1$, equation (4) will always be satisfied provided that

$$| a | + | b | < 1. \tag{5}$$

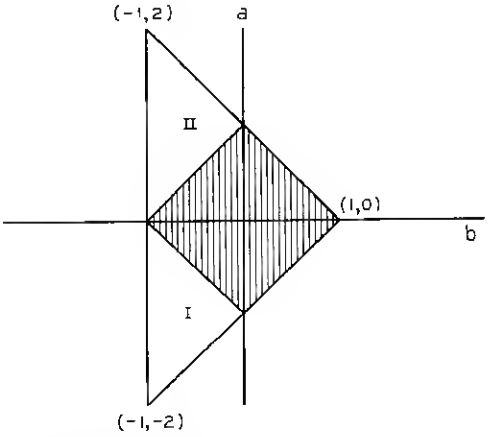

Fig. 3—Some interesting regions in the "space" of feedback tap weights. The hatching indicates stability even with the nonlinearity.

The subset of the a–b plane for which (5) is true is shown in Fig. 3 with vertical hatching, and is a subset of the region of linear stability. It is shown in this Section that if (5) is not satisfied there always exist self-sustained oscillations of the digital filter and hence (5) is hoth a necessary and sufficient condition for absence of self-sustained oscillations.* One way to avoid the oscillations in question is simply to impose the requirement (5). This trick has its limitations, however, for it clearly restricts design capahilities. The region of the s-plane which is shaded in Fig. 4 shows the allowahle pole positions. Roughly speaking, one concludes that there are desirahle filter characteristics that can be realized with this restriction and there are desirable characteristics that cannot.

It is not our purpose here to outline those applications for which (5) will not he restrictive; we proceed to sketch the situation when $| a | + | b | > 1$ and the threat of oscillation is present. Sections III and IV contain, we believe, a novel and interesting mathematical treatment of the general problem of classifying the self-oscillations of the nonlinear difference equation (1). However, for the user of digital filters a simple proof of the $| a | + | b | > 1$ being sufficient for threat of oscillations is of more immediate interest. After reading the simple proof of this fact given next in the present section, such a reader may wish to proceed directly to Section V.

Consider the possibility of undriven nonlinear operation giving a dc

* I. W. Sandberg has informed the authors that the necessity and sufficiency of (5) holding for absence of oscillations has also been obtained jointly by him and L. B. Jackson.

output, that is, $y_k \equiv y$ for all k. Equation (1), with $x(t) = 0$ becomes $y = f[(a + b)y]$. Assuming for definitness that $y > 0$, we can easily see from Fig. 2 that the above equation will be true if $(a + b)y = y - 2$, which implies $y = 2/(1 - a - b)$. One can show (see discussion following equation 17), that this y will have magnitude < 1 provided only that the tap values a and b lie in the region labeled I in Fig. 3. Thus a consistent dc oscillation is always possible for all (a, b) pairs in this region. Next consider the possibility of a period 2 oscillation. This amounts to finding a consistent solution to $y = f[(b - a)y]$. Proceeding as before we obtain

$$y = \frac{2}{1 + a - b}.$$

Thus y_k will be given by $(-1)^k y$, and will have magnitude less than unity if the (a, b) pair lies anywhere in region II of Fig. 3.

III. FURTHER ANALYSIS OF THE OSCILLATIONS

To analyze equation (1) in greater detail, it is very convenient to write it in the form similar to (2),

$$y(t + 2) - ay(t + 1) - by(t) = \sum_n a_n u(t + 2 - n), \qquad (6)$$

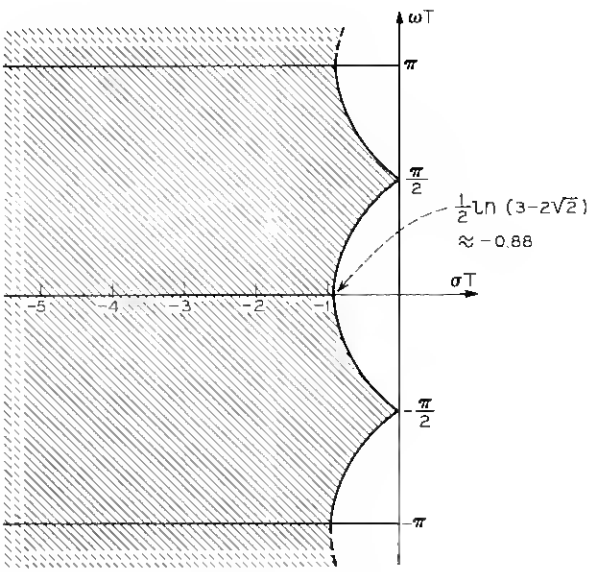

Fig. 4 — Pole locations in the s-plane (shaded region) realizable under the constraint that $|a| + |b| < 1$.

where $u(t)$ is a square pulse of unit height that one may conveniently think of as lasting from $t = 0$ until $t = 1$. This, of course, means that one interprets the solution of (6) to he a piecewise constant function like the actual output of tbe digital filter. For matbematical manipulations it is sometimes desirable to also interpret (6) as a difference equation, defined only for integer t. In this case one would write that $u(t - n) = \delta_{tn}$ where δ_{tn} is the familiar Kroneker symhol.

The point of the right side of (6) is simply to keep $| f(v) | < 1$ regardless of what value v has. From Fig. 2 we see that if $| v | < 1$, this added term is not needed and we take $a_n = 0$. If $1 < v < 3$ then we take $a_n = -2$, and if $-3 < v < -1$ we take $a_n = +2$. Since we have that $| y(t) | < 1$ and that linear stability (see Fig. 3) implies $| a | < 2, | b | < 1$, we need not consider further values of $| v |$. Thus in (6) $a_n = 0, \pm 2$ depending on whetber or not $v(t) \equiv ay(t + 1) + by(t)$ crosses the lines $v = \pm 1$. It will he convenient to have a word for such crossings; we shall call them "clicks", borrowing a favorite word from FM theory. Then $a_n = 0, \pm 2$ depending on whether or not a click does not, or does, occur.

Note if one knew what the click sequence $\{a_n\}$ was, one could solve (6) simply by using the clicks to be tbe driving term for a linear equation. We are mainly interested in describing the self-sustained steady state oscillations of arbitrary period N. Hence initial conditions will play no essential role for us, for while tbey determine whicb oscillating mode appears as $t \to \infty$, they play no role in describing the modes. Our procedure will be as follows:

(*i*) Assume a click sequence of period N;

$$\begin{aligned} &a_0 , a_1 , a_2 , \cdots , a_{N-1} \\ &a_{lN+k} = a_k . \qquad l = 0, 1, \cdots \\ &\qquad\qquad\qquad 0 \leqq k < N - 1. \end{aligned} \tag{7}$$

(*ii*) Using the assumed $\{a_n\}$, find the steady state solution of (6). However, only solutions that have $| y(t) | < 1$ for all t are allowed. (*iii*) Check that this steady state solution actually generates the assumed click sequence.

In carrying out the above program for some simple cases we observed that step *iii* never seemed to yield anything new. Indeed, surprising as it seems at first glance, step *iii* never has to be carried out. If one ohtains a solution with $| y(t) | < 1$, this solution is consistent. That is, it automatically generates the assumed click sequence. The proof is simple.

One calculates the argument of the function f from (6):

$$ay(t+1) + by(t) = y(t+2) - \sum a_n u(t+2-n). \tag{8}$$

We have a click at time $t+2=m$ if $|\, ay(m-2) + by(m-1)\,| > 1$. From (8),

$$|\, ay(m-2) + by(m-1)\,| = |\, y(m) - a_m\,|. \tag{9}$$

Note then if in (9) $a_m = 0$, then $|\, ay(m-2) + by(m-1)\,| = |\, y(m)\,| < 1$; thus if there is no click at a particular time in the assumed click sequence the "solution" will not generate one. Next assume $a_m = +2$; then

$$ay(m-2) + by(m-1) = y(m) - 2 < -1, \tag{10}$$

where we use $|\, y(t)\,| < 1$ again. Equation (10) says if a positive click is present in the assumed click sequence then the solution obtained from the linear equation (6), given by this click sequence, will reproduce the positive click. Obviously the same argument holds for a negative click, $a_m = -2$, and the proof of this point is complete.

The steady-state solution of our fundamental equation (6) for an arbitrary click sequence $\{a_m\}$ of period N is derived in the appendix. If we define

$$A_{N-1}\left(\frac{1}{z}\right) \equiv \sum_{n=0}^{N-1} a_n z^{-n} \tag{11}$$

and

$$D(z) \equiv z^2 - az - b, \tag{12}$$

and let r_i, $i = 1, \cdots, N$, be the N Nth roots of unity, then the (periodic) output values are given by

$$y_k = \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{D(r_i)} r_i^k . \tag{13}$$

The above expression gives the $\{y_k\}$ output sequence for any click sequence. We emphasize, however, that it is only a solution corresponding to a self-sustained oscillation of the digital filter if we have $|\, y_k\,| < 1$, all k. Whether or not this is true depends on the particular click sequence assumed.

Another form of the solution can be obtained by manipulation of (13). To write this down, define

$$b_n^{(k)} \equiv (\bar{a}_{k-1-n} + \bar{a}_{k-1-n+N})/2, \tag{14}$$

where we understand $\bar{a}_i \equiv 0$ if j does not lie between 0 and $N-1$, inclusive, and $\bar{a}_j \equiv a_j$ if it does. One of the $\bar{a}$'s in (14) will thus always be zero and $b_n^{(k)}$ has values of ± 1, 0. The other form of the solution is then

$$y_k = \frac{2}{\rho_1 - \rho_2} \sum_{n=0}^{N-1} b_n^{(k)} \left[\frac{\rho_1^n}{1 - \rho_1^N} - \frac{\rho_2^n}{1 - \rho_2^N} \right]$$

$$k = 0, 1, \cdots, N-1 \qquad (15)$$

where ρ_i are given in (3).

In (15) we have N vectors of dimension N, namely the $\{b_n^{(k)}\}$ $k = 0, 1, 2, \cdots, N-1$. Note from (14), however, that they are all cyclic permutations of one another. Hence we may refer to the b vector, $\mathbf{b}$, of a solution, understanding that the $\mathbf{b}$ and all its cyclic permutations generate a solution in the sense of (15). Note that a cyclic permutation of the y_k has no real significance here; it simply changes the origin of time.

An interesting property of the solutions which we have written down follows from the fact that if we transform the point (a, b) in the ab-plane into another point by

$$\begin{aligned} a &\rightarrow a' = -a \\ b &\rightarrow b' = b \end{aligned} \qquad (16a)$$

then under this transformation

$$\begin{aligned} \rho_1 &\rightarrow \rho_1' = -\rho_2 \\ \rho_2 &\rightarrow \rho_2' = -\rho_1 . \end{aligned} \qquad (16b)$$

The property is this: Let N be an even integer and let $\mathbf{b} = (b_0, b_1, \cdots, b_{N-1})$ be a click vector generating a solution at point (a, b). Then the vector $\mathbf{b}' = (b_0, -b_1, b_2, -b_3, \cdots, b_{N-1})$ generates a solution at reflected point $(-a, b)$. The proof is simple. Note from (15),

$$y'^{(k)} = \frac{2}{\rho_1' - \rho_2'} \sum_n b_n'^{(k)} \left[\frac{\rho_1'^n}{1 - \rho_1'^N} - \frac{\rho_2'^n}{1 - \rho_2'^N} \right]$$

$$= \frac{2}{-\rho_2 + \rho_1} \sum_n (-1)^{k+n} b_n \left[\frac{(-\rho_2)^n}{1 - \rho_2^N} - \frac{(-\rho_1)^n}{1 - \rho_1^N} \right] = (-1)^k y^{(k)} .$$

Hence if $| y^{(k)} | < 1$ then $| y'^{(k)} | < 1$. Note that the proof also supplies the value for $y'^{(k)}$ in terms of $y^{(k)}$. This theorem will be used later to generate new solutions from old ones.

Before leaving this general discussion in favor of exhibiting some solutions in the next section, we list a few more observations related

to the click vector **b**. The click vector **b**, whose only allowed component values are ± 1, 0, completely characterizes the associated oscillation. Clearly there can then only be a finite number of oscillations of given period N. This number is upper bounded by 3^N, but will generally be much less. Also note that a cyclic permutation of the components of b cyclically permutates the output values y^k, and this latter is merely a shift in time. The permutated values are not physically distinct.

Also note that if we perform $\mathbf{b} \rightarrow -\mathbf{b}$ then $\mathbf{y} \rightarrow -\mathbf{y}$, and a solution of opposite sign is obtained. While this may often be distinguishable from the first solution, it is trivially related to it. Finally if one were to count the number **b** vectors of dimension N that yield new information, one would wish to exclude subperiods of N. Thus if (+, 0, 0) is an generating **b** vector for period 3, (+, 0, 0, +, 0, 0) generates a period 6 oscillation but this is not new information. We have not solved the problem of counting how many of the 3^N vectors are left after we impose the requirements of cyclic shifts, sign changes, and subperiods. At any rate, it is essential to test the ones that remain to check that they generate allowed solutions, $| y^k | < 1$.

IV. SOME EXPLICIT PERIODS AND REGIONS OF OSCILLATION

Now for a few explicit solutions. Consider the possibility of a dc "oscillation", namely, set $N = 1$. The only nontrivial click vector is $\mathbf{b} = (+)$. The solution is more immediate if we use (13). We have

$$y = \frac{2}{1 - a - b} \tag{17}$$

for the dc value of output. For what values of a and b within the triangle of Fig. 3 will we have $| y | < 1$? We require

$$| 1 - a - b | > 2 \tag{18}$$

which is equivalent to either

$$1 - a - b > 2 \tag{19a}$$

or

$$-1 + a + b > 2. \tag{19b}$$

Inequality (19a) (coupled with the linear stability requirement) defines the triangle labeled "I" in Fig. 3, while (19b) is outside the stability region and needs no further consideration. Thus any portion of the region $a < 0$ that we have not excluded from oscillations has now been shown to have them. They are of period 1; other period oscillations may (and do) occur in this region.

At this point it is amusing to use an earlier remark on the possibility of generating new solutions from an even period one by "reflection". Letting $N = 2$, the click vector $\mathbf{b} = (+, +)$ certainly generates a period 2 oscillation (albeit one with subperiods) in region I. Then the click vector $\mathbf{b} = (+, -)$ generates something really new: a period 2 oscillation in the region labeled II in Fig. 3. The amplitudes of the output are

$$y^{(k)} = (-1)^k \frac{2}{1 + a - b}, \qquad a > 0. \tag{20}$$

One more possibility of a click vector exists for period 2, and that is $\mathbf{b} = (+, 0)$. From (13) we write for possible output values

$$\begin{aligned} y_0 &= \frac{1}{1 - a - b} + \frac{1}{1 + a - b} \\ y_1 &= \frac{1}{1 - a - b} - \frac{1}{1 + a - b}. \end{aligned} \tag{21}$$

After a little uninteresting analysis one can conclude that we cannot have $| y_0 | < 1, | y_1 | < 1$ in (21) for any allowed values of a and b. Thus there are no other period 2 oscillations.

On to period 3. Now there are four click vectors which must be considered. These are (+00), (++0), (+−0), (++−). Even in this case an exhaustive check that the "solutions" generated are legitimate ones is trying. Therefore, we resort to a trick; we look for periods which may exist in the immediate neighborhood of the point $(a = 0, b = 1)$. This means $\rho_1 = i$, $\rho_2 = -i$. In this immediate neighborhood $\rho_2 = \rho_1^*$, and (15) reads

$$y = \frac{2}{\operatorname{Im} z} \operatorname{Im} \sum_{n=0}^{N-1} \frac{b_n z^n}{1 - z^N}, \tag{22}$$

where we have let $z = \rho_1$. Letting $N = 3$, $z = i$ gives

$$\begin{aligned} y_0 &= -b_0 + b_1 + b_2 \\ y_1 &= -b_1 + b_2 + b_0 \\ y_2 &= -b_2 + b_0 + b_1 . \end{aligned} \tag{23}$$

We now require $y_k = \pm 1$ as a test for the click vector $\mathbf{b}$. We see that only (+00) qualifies as possibly yielding a solution in the neighborhood of $(a = 0, b = -1)$. A computer study shows that indeed the solution extends into the interior of the triangle and the region found is shown in Fig. 5. This immediately implies existence of the period 6 oscillation generated by (+00−00) in the reflected region. Similarly, a period 5 oscillation region (with the concomitant period 10) generated by (+0000) is shown in Fig. 6.

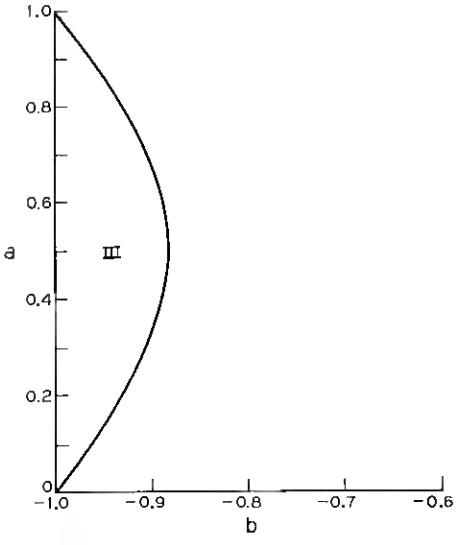

Fig. 5 — A region for period 3 oscillations.

It is very tempting to conjecture that the point ($a = 0, b = -1$) is a boundary point of any allowed region of oscillation. If this is true, a procedure like that used above may eliminate some otherwise very respectable **b** vectors from consideration. Note that for $N = 2$, $b = (+, 0)$ satisfies the required condition at $\rho_1 = i$, but we have shown this

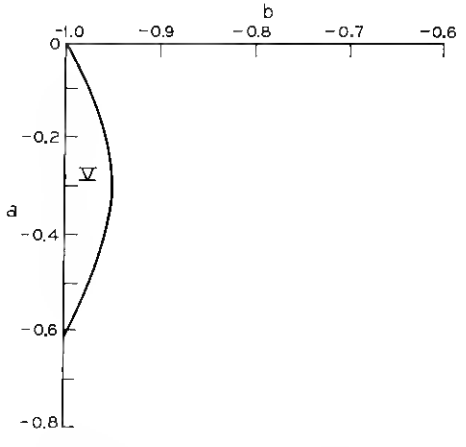

Fig. 6 — A region for period 5 oscillations.

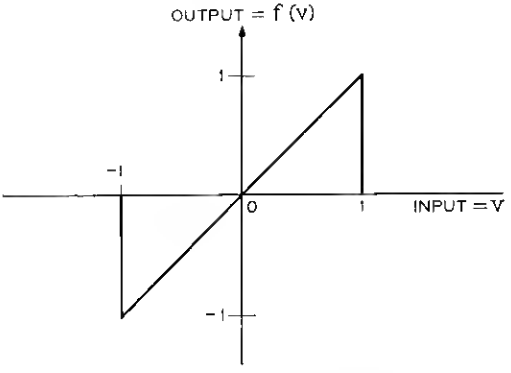

Fig. 7 — Zeroing arithmetic, shown above, also gives rise to oscillations.

is not extendahle into the interior of the triangle. Hence existence at $z = i$ does not guarantee an allowed solution.

V. STABILITY WITH A MODIFIED ARITHMETIC

In an attempt to eliminate these oscillations, proposals have heen made which rely on detecting overflow. One such suggestion dictates that when overflow occurs, the adder is directed to shift out zero. For refereace we call this zeroing arithmetic. The effective transfer function of the adder for zeroing arithmetic is given in Fig. 7. However, it can he shown hy numerical example that such a procedure still leads to oscillations. Another possihility, "saturation arithmetic," is displayed in Fig. 8. Here a one (with the appropriate sign) is put out when overflow is detected. The remaining portion of this paper is devoted to proving that saturation arithmetic leads to stahle operation whenever linear theory would predict it to he so.

To hegin, we suppose for the moment that we ignore the fact that the digitally implemented adder is nonlinear. Then the second-order linear difference equation which governs the hehavior of the undriven system has solutions y_k which may he described as follows:

Case 1: Complex roots for characteristic equation

$$y_k = \text{Re } K_0 \exp(-\alpha k), \quad K_0 \text{ and } \alpha \text{ complex}, \quad \text{Re } \alpha > 0. \qquad k = 0, 1, 2, \cdots. \tag{24}$$

Case 2: Real but unequal roots

$$y_k = K_1 \exp(-\alpha k) + K_2 \exp(-\beta k). \quad K_i \text{ real}; \quad \alpha > 0, \quad \beta > 0. \tag{25}$$

Case 3: Real and equal roots

$$y_k = [K_1 + K_2 k] \exp(-\alpha k). \quad K_i \text{ real}; \quad \alpha > 0. \tag{26}$$

Using this information, coupled with knowledge of y_j and y_{j+1} for some j, it is easy to give a bound on the magnitudes of all future ($k \geqq j$) values of the output and to show this value goes to zero with increasing j. This is just another way to say that the solutions go to zero for the linear case. In the nonlinear case we cannot exclude the situation that some y_{k+1} will exceed unity and the nonlinearity will he operative. For saturation arithmetic the offending value must he set to unity if, for example, $y_{k+1} > +1$. We can, for conceptual purposes, regard this as a "squeezing" of the output from a value greater than unity down to the value one which is performed in a continuous fashion. The crux of the proof now comes in showing that the partial derivative of our hound (on future outputs) with respect to the most recent output y_{k+1} has, for saturation arithmetic, the same sign as y_{k+1}. Hence decreasing a value that is too large in magnitude will decrease the bound as well, and it will go to zero at least as fast as it does for the linear case.

To show how the above outline works, consider first the linear case with complex roots. From the form of the solution

$$y_k = \text{Re } K_0 \exp(-\alpha k), \quad \text{Re } \alpha > 0, \quad k = 0, 1, 2, \cdots,$$

it is clear that if we define

$$B_0 = |K_0|^2 \tag{27}$$

then $y_k^2 \leqq B_0$ for all $k \geqq 0$. We now express B_0 in terms of the values y_0, y_1 which are initially stored in the shift registers to yield

$$B_0 = y_0^2 + \frac{[y_1 - y_0 \text{ Re exp } (-\alpha)]^2}{[\text{Im exp } (-\alpha)]^2}. \tag{28}$$

This suggests that one define the more general set of numbers

$$B_i = y_i^2 + \frac{[y_{i+1} - y_i \text{ Re exp } (-\alpha)]^2}{[\text{Im exp } (-\alpha)]^2}. \tag{29}$$

Clearly, from the way that B_j is defined, we have tbat

$$y_k = \text{Re } K_j \exp[-\alpha(k - j)], \qquad k \geqq j \tag{30}$$

where K_j is some appropriate complex number that satisfies

$$B_j = |K_j|^2. \tag{31}$$

From (30), the additional inequality that $y_k^2 \leqq B_j$ for all $k \geqq j$ follows.

Furthermore, one can see by comparing (30) and (24) that

$$| K_j |^2 = | K_0 |^2 | \exp(-\alpha j) |^2. \tag{32}$$

Hence, since the real part of α is positive, B_j goes monotonically to zero with increasing j.

To generalize the above arguments to a nonlinear situation of interest,* consider the following equation which follows from (29):

$$\frac{\partial B_j}{\partial y_{j+1}} = \frac{2}{[\text{Im } \exp(-\alpha)]^2} [y_{j+1} - y_j \text{ Re } \exp(-\alpha)]. \tag{33}$$

Now imagine B_{j-1} has been calculated from values stored in the registers. From *linear* theory we predict $y_{j+1}^{(L)}$ and $B_j^{(L)} \leqq B_{j-1} \exp(-2\alpha)$, by (32). Now if the $y_{j+1}^{(L)}$ generated by the linear equation were too large, say, then decreasing it to unity would, according to (33), *decrease* the bound B_j if we knew that

$$y_{j+1} - y_j \text{ Re } [\exp(-\alpha)] \geqq 0 \quad \text{for} \quad y_{j+1}^{(L)} \geqq y_{j+1} \geqq y_{j+1}^{(C)} \tag{34}$$

where $y_{j+1}^{(L)}$ is the linear prediction for y_{j+1} and $y_{j+1}^{(C)}$ is the correct value for the nonlinear circuit resulting from "squeezing" $y_{j+1}^{(L)}$ down. Since $| y_j | \leqq 1$ and Re $\exp(-\alpha) < 1$, (34) is always true for saturation arithmetic (see Fig. 8) because $y_{j+1}^{(C)} = +1$ (assuming $y_{j+1}^{(L)} > +1$) and (34) can never swing negative. Similar things happen, of course, if $y_{j+1} < -1$. Thus the bound decreases at least as fast as for the linear case (which is exponential) and stability is assured. For zeroing arithmetic $y_{j+1}^{(C)} = 0$, and thus the appropriate sign for (34) cannot be guaranteed which is in satisfying agreement with the known instability for this case.

For the next case of real but unequal roots, we now have reference to equation (25) and define our initial bound as

$$B_0 = 2(K_1^2 + K_2^2)$$

$$= 2 \frac{[y_1 - \exp(-\alpha) y_0]^2 + [y_1 - \exp(-\beta) y_0]^2}{[\exp(-\alpha) - \exp(-\beta)]^2}. \tag{35}$$

The remaining details are too similar to those of the preceding case to warrant recording again; stability for saturation arithmetic holds here as well.

The last case to discuss occurs when we have real and equal roots.

---

* B_j calculated from (29) is a bound on future outputs for the nonlinear as well as the linear case. If $B_j \leq 1$ the two cases coincide, while of $B_j > 1$ the conclusion follows equally trivially since $|y_k| \leq 1$ for the nonlinear situation.

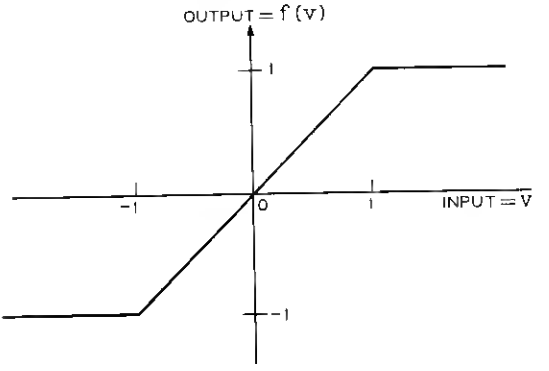

Fig. 8 — The above nonlinearity corresponds to saturation arithmetic and leads to stable behavior.

This situation, represented for the linear equation by equation (26), is more difficult to treat than the previous ones. The analog of (27) and (35) now is

$$B_0 = \max \begin{cases} 4K_1^2 \\ \dfrac{4K_2^2}{\alpha^2} \end{cases}. \tag{36}$$

That (36) yields a bound follows from the facts that (for $t \geqq 0$)

$$\begin{aligned} y_k^2 &\leqq \max_t [(K_1 + K_2 t) \exp(-\alpha t)]^2 \\ &\leqq 2 \max_t [K_1^2 + K_2^2 t^2] \exp(-2\alpha t) \\ &\leqq 4 \max \begin{cases} \max_t K_1^2 \exp(-2\alpha t) \\ \max_t K_2^2 t^2 \exp(-2\alpha t) \end{cases} \\ &= 4 \max \begin{cases} K_1^2 \\ \dfrac{K_2^2 \exp(-2)}{\alpha^2} \end{cases} \\ &\leqq 4 \max \begin{cases} K_1^2 \\ \dfrac{K_2^2}{\alpha^2} \end{cases}. \end{aligned}$$

Since

$$\begin{aligned} K_1^2 &= y_0^2 \\ \frac{K_2^2}{\alpha^2} &= \frac{(y_1 \exp \alpha - y_0)^2}{\alpha^2}, \end{aligned} \tag{37}$$

we define our general bound as

$$B_j = 4 \max \begin{cases} y_j^2 \\ \dfrac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2}. \end{cases} \tag{38}$$

Using the solution $y_j = (K_1 + K_2 j) \exp(-\alpha j)$, we see that

$$\theta_j \equiv \frac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2} \tag{39}$$

decreases by the multiplicative factor $\exp(-2\alpha)$ for every unit increase of j. Further, suppose that $B_j = 4y_j^2$ for some j. That is, suppose

$$\frac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2} < y_j^2 . \tag{40}$$

This implies

$$y_{j+1}^2 < y_j^2 (1 + \alpha)^2 \exp(-2\alpha), \tag{41}$$

and so if next time $B_{j+1} = 4y_{j+1}^2$, then we have decreased by $(1 + \alpha)^2 \exp(-2\alpha) < 1$. On the other hand, if at the next step we have to choose $B_{j+1} = 4\theta_{j+1}$, we see

$$\frac{B_{j+1}}{B_j} = \frac{\theta_{j+1}}{y_j^2} \leqq \frac{\theta_{j+1}}{\theta_j} \leqq \exp(-2\alpha). \tag{42}$$

Likewise if we go from $4\theta_j$ to $4\theta_{j+1}$ we decrease by $\exp(-2\alpha)$. Finally, a "transition" from $4\theta_j$ as a bound to $4y_{j+1}^2$ decreases the bound by a multiplicative factor of $(1 + \alpha)^2 \exp(-2\alpha)$. To see this we note that, by assumption,

$$B_j = \frac{4[y_{j+1} \exp \alpha - y_j]^2}{\alpha^2} \geqq 4y_j^2 . \tag{43}$$

Using the left-hand equality in (43) implies

$$| y_{j+1} | \exp \alpha \leqq \frac{\alpha (B_j)^{\frac{1}{2}}}{2} + | y_j |. \tag{44}$$

while $B_i \geqq 4y_i^2$ yields

$$| y_i | \leqq \frac{(B_i)^{\frac{1}{2}}}{2}. \tag{45}$$

Using (45) in (44) then allows us to deduce that

$$B_{i+1} = 4y_{i+1}^2 \leqq (1 + \alpha)^2 \exp(-2\alpha)B_i \tag{46}$$

as was claimed. To extend these arguments to the nonlinear case we again observe that

$$\frac{\partial B_i}{\partial y_{i+1}} \geqq 0 \tag{47}$$

for saturation arithmetic.

VI. GENERALIZATIONS TO OTHER STABLE NONLINEARITIES

Aside from the three nonlinearities already mentioned, there does not appear to be immediate engineering interest in seeing which other nonlinearities will or will not give rise to stable behavior of the filter. Having come this far, however, it is hard to resist asking if the method of proof we have used, or some slight extension of it, does suggest other nonlinearities for which stability will hold. The extension we consider is not to require

$$\frac{\partial B_i}{\partial y^{i+1}} \geqq 0$$

all during the "squeezing" operation, but merely that

$$B_i^L - B_i^C \geqq 0, \tag{48}$$

where B_i^L is the value of the bound using linear theory and B_i^C is the "correct" value. An inspection of the previous proofs shows that this is equivalent to

$$(y_{i+1}^L - ay_i)^2 - (y_{i+1}^C - ay_i)^2 > 0 \tag{49}$$

for all real a such that $| a | < 1$.

A little manipulation reduces (49) to

$$(y_{k+1}^L - y_{k+1}^C)(y_{k+1}^L + y_{k+1}^C - 2ay_k) \geqq 0. \tag{50}$$

Assuming $y_{k+1}^L > 0$, the first term in (50) to be nonnegative, and $| y_k | \leqq 1$, makes it apparent that

$$y_{k+1}^L + y_{k+1}^C \geqq 2 \tag{51}$$

is sufficient. The "stable nonlinearities" deduced from this kind of reasoning are outlined in Fig. 9. Thus any nonlinearity whose graph coincides with the identity function on the interval $[-1, 1]$ and whose remaining portions lie in the closed shaded region of Fig. 9 will he stahle. The function in these regions need not be continuous and need not ohey $f(-u) = -f(u)$.

An even higher degree of generality is achieved when we realize tbat nothing in our proofs required the nonlinearity $f(u)$ to be tbe same for successive values of the parameter k. This is tantamount to allowing the nonlinearity to be random in the following manner. Suppose a value of $y_{k+1}^L > 1$ has heen predicted from linear theory (see Fig. 9). The perpendicular P to the v axis through y_{k+1}^L intersects the sbaded region shown in Fig. 9 along a line segment. Cboose randomly from this line segment the "value" of tbe nonlinearity to give y_{k+1}^C. The discussion in this Section shows that the solutions of the difference equation

$$y_{k+2} = f[ay_{k+1} + by_k] \tag{52}$$

which has the stochastic nonlinearity just descrihed will be stahle wbenever the linear version has stahle solutions.

APPENDIX

Derivation of the Steady-State Solution

We obtain the steady-state solution of our fundamental equation (6) using z-transforms. Recall that if one has a hounded sequence of number $\{a_n\}$, the z-transform is defined by

$$f(z) = \sum_{n=0}^{\infty} a_n z^{-n} \tag{53}$$

where (53) converges and is analytic outside the unit circle, $|z| > 1$. It is easy to sbow tbat if $\{a_n\}$ is periodic of period N, that is if $a_{N+n} = a_n$, then (53) becomes

$$f(z) = \frac{A_{N-1}\left(\frac{1}{z}\right)}{1 - z^{-N}} \tag{54}$$

where A_{N-1} is the polynomial of degree $(N - 1)$ in $1/z$ given hy

$$A_{N-1}\left(\frac{1}{z}\right) = \sum_{n=0}^{N-1} a_n z^{-n}. \tag{55}$$

The N poles of $f(z)$ at the N roots of unity are apparent from (12), and there are no other poles.

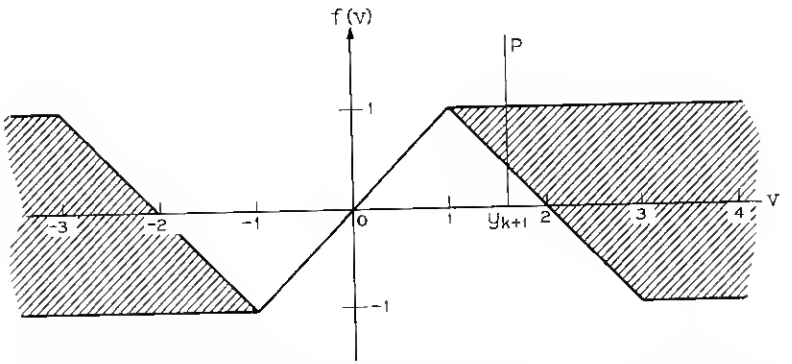

Fig. 9 — Any nonlinearity whose graph coincides with the identity function on the interval $[-1, +1]$ and whose remaining portions lie in the (closed) shaded region will be stable. The possibility of generalizing this to a stochastic nonlinearity is also noted in the text.

Denoting by $Y(z)$ the z-transform of $y(t)$ *excluding* the additive terms involving initial conditions (since these will damp out because of linear stability) we have from (6) that

$$Y(z) = \frac{A_{N-1}\left(\frac{1}{z}\right)}{(z^2 - az - b)(1 - z^{-N})}. \tag{56}$$

The z-transform of the steady-state solution $\hat{Y}(z)$ must still be extracted from $Y(z)$. Since the unit circle $|z| = 1$ corresponds to the frequency axis if one were using Fourier transforms, we know, by analogy, the state steady-state portion of (56) will be the pole-terms. Let r_i, $i = 1, \cdots, N$ be the N Nth roots of unity and define

$$Q_i^{N-1}\left(\frac{1}{z}\right) \equiv \sum_{k=0}^{N-1} \left(\frac{1}{r_i}\right)^{N-1-k}\left(\frac{1}{z}\right)^k = \frac{1 - z^{-N}}{\frac{1}{r_i} - \frac{1}{z}}. \tag{57}$$

Note (57) implies

$$Q_i^{N-1}\left(\frac{1}{r_i}\right) = Nr_i\ . \tag{58}$$

Then from (56)–(58) we have

$$\hat{Y}(z) = \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{\left(\frac{1}{r_i} - \frac{1}{z}\right) \cdot Nr_i \cdot D(r_i)}\ , \tag{59}$$

where we have let

$$D(z) = z^2 - az - b. \tag{60}$$

Using (57) once more, the steady-state solution (59) may be written

$$\hat{Y}(z) = \frac{1}{1 - z^{-N}} \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right) Q_i^{N-1}\left(\frac{1}{z}\right)}{r_i D(r_i)}. \tag{61}$$

Referring back to the discussion at the beginning of this section, we see that (61) is the z-transform of a sequence $\{y_k\}$ of period N where

$$y_k = \text{coefficient of } z^{-k} \text{ in } \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right) Q_i^{N-1}\left(\frac{1}{z}\right)}{r_i D(r_i)} \right\}$$

$$k = 0, 1, \cdots, N - 1. \tag{62}$$

Using (57) in (62) we obtain

$$y_k = \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{D(r_i)} r_i^k, \tag{63}$$

where, in writing (63), we have used the fact that $r_i^N = 1$. Expression (63) thus gives the $\{y_k\}$ sequence for any click sequence. It is a solution corresponding to a self-sustained oscillation of the digital filter only if we have $| y_k | < 1$, all k.

Two sums appear in (63). The explicit one shown is the sum over the roots of unity; the hidden one is the polynomial $A_{N-1}(1/r_i)$. We will exhibit another form of solution (63) by explicitly doing the sum over the N roots. We begin by writing

$$A_{N-1}\left(\frac{1}{r_i}\right) = 2 \sum_{l=0}^{N-1} \frac{p_l}{r_i^l}, \qquad p_l = \pm 1, 0. \tag{64}$$

Thus p_l are the coefficients, except for the factor of 2, of the polynomial $A_{N-1}(z)$. We also write, by factoring $D(z)$ and expanding in partial fractions,

$$\frac{1}{D(z)} = \frac{1}{(z - \rho_1)(z - \rho_2)} = \frac{1}{\rho_1 - \rho_2} \left[\frac{1}{z - \rho_1} - \frac{1}{z - \rho_2} \right]. \tag{65}$$

Now note that if z is such a number than $z^N = 1$, we have (since $| \rho | < 1$ and $| z | = 1$)

$$\frac{1}{z-\rho} = \frac{1}{z}\sum_{n=0}^{\infty}\left(\frac{\rho}{z}\right)^n. \tag{66}$$

Let us look at the sum of the $n = 0, N, 2N$, etc., terms in the right side of (66), that is

$$1 + \frac{\rho^N}{z^N} + \frac{\rho^{2N}}{z^{2N}} + \frac{\rho^{3N}}{z^{3N}} + \cdots$$

$$= 1 + \rho^N + \rho^{2N} + \rho^{3N} + \cdots = \frac{1}{1-\rho^N}. \tag{67}$$

Treating the sum of terms

$$n = 1, N+1, 2N+1, \cdots$$
$$n = 2, N+2, 2N+2, \cdots$$
$$\vdots$$
$$n = N-1, N+(N-1), 2N+(N-1), \cdots$$

similarly, we have

$$\frac{1}{z-\rho} = \frac{1}{z}\cdot\frac{1}{1-\rho^N}\left[1 + \frac{\rho}{z} + \frac{\rho^2}{z^2} + \cdots + \frac{\rho^{N-1}}{z^{N-1}}\right]. \tag{68}$$

Finally letting $z = 1/r_i$ gives

$$\frac{1}{\frac{1}{r_i}-\rho} = \frac{r_i}{1-\rho^N}\sum_{n=0}^{N-1}[\rho r_i]^n. \tag{69}$$

Using (65) and (64) in (63) yields

$$y_k = \frac{1}{\rho_1-\rho_2}\cdot\frac{2}{N}\sum_i r_i^k\left(\sum_{l=0}^{N-1}\frac{p_l}{r_i^l}\right)$$
$$\cdot\left[\frac{1}{r_i}\sum_{n=0}^{N-1}\frac{1}{r_i^n}\left(\frac{\rho_1^n}{1-\rho_1^n} - \frac{\rho_2^n}{1-\rho_2^n}\right)\right]. \tag{70}$$

Two sums in (70) are immediately done. First look at the sum over the roots of unity. This involves observing that

$$\sum_i r_i^{k-l-1-n} = \begin{cases} N & \text{if} \quad k-l-1-n \equiv 0 \mod N, \\ 0 & \text{otherwise.} \end{cases} \tag{71}$$

The congruence indicated in (71) can only be satisfied here if $l = k -$

$1 - n$ or if $l = k - 1 - n + N$. Thus it is useful to define

$$2b_n^{(k)} \equiv \bar{a}_{k-1-n} + \bar{a}_{k-1-n+N}, \tag{72}$$

where we understand $\bar{a}_j \equiv 0$ if j does not lie between 0 and $N - 1$, inclusive, and $\bar{a}_j \equiv a_j$ if it does. One of the $\bar{a}$'s in (72) will thus always be zero and $b_n^{(k)}$ has values, like the p's, of $\pm 1, 0$. Using the discussion above surrounding equations (71) and (72) we perform next the sum over l and write another form of the solution:

$$y_k = \frac{2}{\rho_1 - \rho_2} \sum_{n=0}^{N-1} b_n^{(k)} \left[\frac{\rho_1^n}{1 - \rho_1^N} - \frac{\rho_2^n}{1 - \rho_2^N} \right]$$

$$k = 0, 1, \cdots, N - 1. \tag{73}$$

REFERENCES

1. Rader, C. M., Gold, B., "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55*, No. 2 (February 1967), pp. 149–171.
2. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio and Electroacoustics, *AV-16*, No. 3 (September 1968), pp. 413–421.